MEASUREMENT-BASED CONSTRUCTION OF LOCALITY-AWARE OVERLAY NETWORKS

FIELD OF THE INVENTION

[0001] This invention pertains generally to computer networks and, more particularly, to computer networks capable of supporting one or more overlay networks.

BACKGROUND OF THE INVENTION

[0002] The number of applications that take advantage of services offered by computer networks is large and growing. Such applications include interactive multimedia communication, multimedia file retrieval, streaming media distribution, and distributed computing. A basic service typically provided by a modern computer network is the ability for each computer in the network to establish a communication connection with any other computer in the network. Particularly in large computer networks, it is common for this service to be provided by sophisticated routing rather than, for example, brute each-to-each (NxN) connectivity. For at least the purposes of this description, the computer network involved in providing this basic service is called the transport network.

[0003] Various application architectures may utilize the transport network in different ways. For example, a client-server architecture may have a centralized server component and a number of client components, located throughout the transport network, that establish more or less temporary communication connections with the server as required. Another example is the

recently popular peer-to-peer (P2P) architecture. Peer-to-peer architectures typically avoid centralized elements, instead relying on peers with roughly equivalent functionality to provide the backbone of the application. The benefits of such decentralization may include avoiding single points of failure and vulnerability to attack, as well as good theoretical scalability.

[0004] There are various conventional schemes for arranging peers in a peer-to-peer architecture. Examples of such schemes include the CAN architecture described by Ratnasamy et al. in *A Scalable Content-Addressable Network*, Proceedings of ACM SIGCOMM, August 2001, the Chord architecture described by Stoica et al. in *Chord: A scalable Peer-to-peer Lookup Service for Internet Applications*, Proceedings of ACM SIGCOMM, August 2001, the Pastry architecture described by Rowstron et al. in *Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems*, Proceedings of the 18th IFIP/ACM International Conference on Distributed Systems Platforms (Middleware 2001), November 2001, and the Tapestry architecture described by Zhao et al. in *Tapestry: An Infrastructure for Fault-tolerant Wide-area Location and Routing*, Report No. UCB/CSD-01-1141, Computer Science Division, University of California, Berkeley, April 2001. In general, peers in a peer-to-peer architecture are arranged in a network, called an overlay network, which may be considered independently of the underlying transport network.

[0005] A problem with constructing the overlay network of a peer-to-peer architecture independently of the transport network is that peers that are neighbors in the overlay network may be far from each other in the underlying transport network and so, for example, may experience long delays when communicating with each other. This discord between the overlay network and the

transport network is not wholly undesirable because, for example, it may enable the overlay network, and thus an application utilizing it, to tolerate failures in the transport network. However, many applications that desire the advantages of overlay networks, particularly those with high quality of service (QoS) requirements, would benefit from a system and method for constructing overlay networks that make efficient use of the underlying transport network.

BRIEF SUMMARY OF THE INVENTION

[0006] This section presents a simplified summary of some embodiments of the invention. This summary is not an extensive overview of the invention. It is not intended to identify key/critical elements of the invention or to delineate the scope of the invention. Its sole purpose is to present some embodiments of the invention in a simplified form as a prelude to the more detailed description that is presented later.

[0007] In an embodiment of the invention, overlay network peers are grouped so that each peer in a peer group has a similar transport network proximity measure with respect to the peers in other peer groups.

[0008] A first set of transport network distances may include transport network distances between an overlay network peer group and peer group neighbors of the overlay network peer group. A second set of transport network distances may include transport network distances between a peer and the peer group neighbors of the overlay network peer group. In an embodiment of the invention, the peer decides to join the overlay network peer group if the first set of transport network distances is near to the second set of transport network distances. In an embodiment

of the invention, a join locality-aware overlay module is configured to determine if the first set of transport network distances is near to the second set of transport network distances.

[0009] In an embodiment of the invention, a first overlay network peer group queries a second overlay network peer group for the second peer group's neighboring peer groups. Each overlay network peer group has at least one neighboring peer group unless it is the first peer group in an overlay network. The transport network distance between the first overlay network peer group and each of the second peer group's neighbors is measured. At least one overlay network connection is established between the first overlay network peer group and the closest of the second peer group's neighbors.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] While the appended claims set forth the features of the invention with particularity, the invention and its advantages are best understood from the following detailed description taken in conjunction with the accompanying drawings, of which:

[0011] Figure 1 is a schematic diagram illustrating computers connected by a transport network.

[0012] Figure 2 is a schematic diagram generally illustrating an example computer system usable to implement an embodiment of the invention;

[0013] Figure 3 is a schematic diagram of an example overlay network constructed independently of its underlying transport network;

[0014] Figure 4 is a schematic diagram of the example overlay network of Figure 3 in which the illustrated distances between the peers are representative of the transport network distances between the peers;

[0015] Figure 5 is a schematic diagram of an example locality-aware overlay network in accordance with an embodiment of the invention;

[0016] Figure 6 is a block diagram of an example modular software architecture suitable for implementing a locality-aware peer in accordance with an embodiment of the invention;

[0017] Figure 7 is a schematic diagram of peer group network aspects of an example locality-aware overlay network in accordance with an embodiment of the invention;

[0018] Figure 8 is a schematic diagram of network layers associated with a conventional overlay network;

[0019] Figure 9 is a schematic diagram of network layers associated with a locality-aware overlay network in accordance with an embodiment of the invention;

[0020] Figure 10 is a flowchart depicting example steps performed by a locality-aware peer when joining a locality-aware overlay network in accordance with an embodiment of the invention; and

[0021] Figure 11 is a flowchart depicting example steps performed by a locality-aware peer when establishing a new peer group in a locality-aware overlay network in accordance with an embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

[0022] Prior to proceeding with a description of the various embodiments of the invention, a description of a computer and

networking environment in which the various embodiments of the invention may be practiced is now provided. Although not required, the invention will be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, programs include routines, objects, components, data structures and the like that perform particular tasks or implement particular abstract data types. The term "program" as used herein may connote a single program module or multiple program modules acting in concert. The terms "computer" and "computing device" as used herein include any device that electronically executes one or more programs, such as personal computers (PCs), hand-held devices, multi-processor systems, microprocessor-based programmable consumer electronics, network PCs, minicomputers, tablet PCs, laptop computers, consumer appliances having a microprocessor or microcontroller, routers, gateways, hubs and the like. The invention may also be employed in distributed computing environments, where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, programs may be located in both local and remote memory storage devices.

[0023] An example of a computer networking environment suitable for incorporating aspects of the invention is described with reference to Figure 1. The example computer networking environment 100 includes several computers 102 communicating with one another over a network 104, represented by a cloud. Network 104 may include many well-known components, such as routers, gateways, hubs, etc. and allows the computers 102 to communicate via wired and/or wireless media. When interacting with one another over the network 104, one or more of the computers 102 may act as clients, servers or

peers with respect to other computers 102. Accordingly, the various embodiments of the invention may be practiced on clients, servers, peers or combinations thereof, even though specific examples contained herein may not refer to all of these types of computers. Computer networking environment 100 is an example of a transport network.

[0024] Referring to Figure 2, an example of a basic configuration for the computer 102 on which aspects of the invention described herein may be implemented is shown. In its most basic configuration, the computer 102 typically includes at least one processing unit 202 and memory 204. The processing unit 202 executes instructions to carry out tasks in accordance with various embodiments of the invention. In carrying out such tasks, the processing unit 202 may transmit electronic signals to other parts of the computer 102 and to devices outside of the computer 102 to cause some result. Depending on the exact configuration and type of the computer 102, the memory 204 may be volatile (such as RAM), non-volatile (such as ROM or flash memory) or some combination of the two. This most basic configuration is illustrated in Figure 2 by dashed line 206.

[0025] The computer 102 may also have additional features/functionality. For example, computer 102 may also include additional storage (removable 208 and/or non-removable 210) including, but not limited to, magnetic or optical disks or tape. Computer storage media includes volatile and non-volatile, removable and non-removable media implemented in any method or technology for storage of information, including computer-executable instructions, data structures, program modules, or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory, CD-ROM,

digital versatile disk (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to stored the desired information and which can be accessed by the computer 102. Any such computer storage media may be part of computer 102.

[0026] The computer 102 preferably also contains communications connections 212 that allow the device to communicate with other devices such as remote computer(s) 214. A communication connection is an example of a communication medium. Communication media typically embody computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. By way of example, and not limitation, the term "communication media" includes wireless media such as acoustic, RF, infrared and other wireless media. The term "computer-readable medium" as used herein includes both computer storage media and communication media.

[0027] The computer 102 may also have input devices 216 such as a keyboard/keypad, mouse, pen, voice input device, touch input device, etc. Output devices 218 such as a display, speakers, a printer, etc. may also be included. All these devices are well known in the art and need not be described at length here.

[0028] In the description that follows, the invention will be described with reference to acts and symbolic representations of operations that are performed by one or more computing devices, unless indicated otherwise. As such, it will be understood that such acts and operations, which are at times referred to as being computer-executed, include the

manipulation by the processing unit of the computer of electrical signals representing data in a structured form. This manipulation transforms the data or maintains it at locations in the memory system of the computer, which reconfigures or otherwise alters the operation of the computer in a manner well understood by those skilled in the art. The data structures where data is maintained are physical locations of the memory that have particular properties defined by the format of the data. However, while the invention is being described in the foregoing context, it is not meant to be limiting as those of skill in the art will appreciate that various of the acts and operation described hereinafter may also be implemented in hardware.

[0029] Proximity in a communications network may be measured in terms of time. Suitable proximity metrics/measures include send-receive latency, round trip time and network routing hops, so that, for example, two points in the network may be said to be close if the round trip time between them is low or far if the number of hops is high. There are at least two such distances associated with each pair of peers in an overlay network: a first distance as measured through the overlay network and a second distance as measured through the underlying transport network. The metrics may have the same value, for example, if each metric is round trip time, or they may differ, for example, a direct connection may exist between two peers in the overlay network when five routing hops are required through the transport network. Unless stated otherwise, or clearly contradicted by context, references to the distance between two peers in the description below will refer to the distance as measured through the underlying transport network. In an embodiment of the invention, the

overlay network is constructed so as to minimize the average transport network distance between neighboring peers.

[0030] Figure 3 illustrates an example overlay network 300 constructed independently of its underlying transport network. Accordingly, the illustrated distances between peers in Figure 3 are independent of the underlying transport network distances between the peers. Such overlay networks are known in the art, so only some of their details are described here. The computer networking environment 100 discussed with reference to Figure 1 is an example of a suitable transport network. The computers 102 of the transport network host the peers 302, 304, 306, 308, 310, 312, 314, 316 of the overlay network 300. Both computers and peers may be considered as nodes in the transport network. The relationship between a peer (e.g., the peer 302) and a computer (e.g., the computer 102) is not necessarily one-to-one. A single computer may host a plurality of peers, for example, peer 302 and peer 316 may be hosted by a single computer 102. A single peer (e.g., peer 306) may be hosted by a plurality of computers 102.

[0031] The peer interconnection details of overlay network 300 are relatively unimportant as the illustration is intended to be schematic of any overlay network constructed independently of its underlying transport network, however, they are described here for completeness. Peer 302 is directly connected to (is a neighbor in the overlay network of) peers 304, 306 and 316. Peer 304 is connected to peers 302 and 306. Peer 306 is connected to peers 302, 304, 308 and 310. Peer 308 is connected to peers 306 and 310. Peer 310 is connected to peers 306, 308 and 312. Peer 312 is connected to peers 310, 314 and 316. Peer 314 is connected to peers 312 and 316. Peer 316 is connected to peers 302, 312 and 314. Peers connected in

this sense need not have an associated active transport network communication connection (e.g., communication connection 212 of Figure 2), but each such connection typically does have some transport network and/or overlay network resources allocated to it.

[0032] Figure 4 shows the same overlay network 300 as Figure 3 except that the illustrated distances between peers in Figure 4 are representative of the distances (i.e., the proximity metrics) between the peers in the underlying transport network. Peers 302, 306, 312 and 316 are close together in the transport network. They are within a first transport network locality delimited by dashed line 402 and together form peer group A. Peers 304, 308, 310 and 314 are also close together in the transport network. They are within a second transport network locality delimited by dashed line 404 and form peer group B. Peer group A and peer group B are relatively far from each other in the transport network, that is, in this example, connections between peers in different groups are assumed to be long distance connections.

[0033] Peer 302 and peer 316 are neighbors in the overlay network 300. Figure 4 shows that peer 302 and peer 316 are also close together in the transport network. Peer 304 and peer 314 are close together in the transport network, but overlay network messages from peer 304 to peer 314 routed through even the shortest path in the overlay network 300 are first sent from peer 304 to peer 302, and then from peer 302 to peer 316, and finally from peer 316 to peer 314. The overlay network 300 connections from peer 304 to peer 302 and from peer 316 to peer 314 are long distance connections. Even though peer 304 and peer 314 are close together in the transport network, overlay network messages routed between them incur the

cost of two long distance connections. This is an example of an inefficiency targeted by an embodiment of the invention.

[0034] Figure 5 shows the same peers and transport network localities as Figure 4, but in Figure 5, the example peer-to-peer overlay network 500 has been constructed with an awareness of the underlying transport network localities 402 and 404, that is, the overlay network 500 is locality-aware in accordance with an embodiment of the invention. In comparison with the overlay network 300 of Figure 3 and Figure 4, the overlay network 500 has fewer long distance (i.e., inter-group) connections. In an embodiment of the invention, the average transport network distance between neighboring peers in the overlay network is reduced by grouping the peers of the overlay network in transport network localities and minimizing the number of inter-group connections.

[0035] In keeping with the decentralized nature of overlay networks, it is desirable that each peer of a locality-aware overlay network in accordance with an embodiment of the invention is able to detect transport network localities and self-organize into interconnected groups. Figure 6 illustrates an example modular software architecture for a locality-aware peer 600. The locality-aware peer 600 has a join locality-aware overlay module 602 which enables the locality-aware peer 600 to join a locality-aware overlay network peer group in its transport network locality. An intra-group maintenance module 604 maintains an intra-group cache 606. The intra-group cache 606 includes information regarding the peer group in which the locality-aware peer 600 participates. An inter-group maintenance module 608 maintains an inter-group cache 610. The inter-group cache 610 includes information regarding peer

groups other than the peer group in which the locality-aware peer 600 participates.

[0036] A rendezvous point module 612 maintains and provides access to a boot peer cache 614. The boot peer cache 614 includes a list of one or more peers in the overlay network that may be utilized as starting points for the locality-aware peer 600 when it is searching for a peer group in its transport network locality. The dashed line 616 indicates that, in various embodiments of the invention, the rendezvous point module 612 and the boot peer cache 614 may be incorporated by all, some or none of the peers in the locality-aware overlay network. The rendezvous point module 612 and the boot peer cache 614 may be incorporated into a peer (not necessarily the locality-aware peer 600) at a well-known (or easily discoverable) transport network location that is dedicated to providing rendezvous point functionality. Lightweight Directory Access Protocol (LDAP) and Dynamic Host Configuration Protocol (DHCP) servers and the like may be utilized to implement rendezvous point functionality. Distributing multiple boot peers (i.e., peers incorporating the rendezvous point module 612 and the boot peer cache 614) throughout the overlay network may enhance the robustness and fault-tolerance of the overlay network.

[0037] In an embodiment of the invention, the locality-aware peer 600 joining the locality-aware overlay network (e.g., the locality-aware overlay network 500 of Figure 5 or the locality-aware overlay network 700 described below with reference to Figure 7) either joins a peer group in the transport network locality of the peer 600 or, if the peer 600 fails to find a peer group in its locality, the peer 600 establishes a new peer group. When joining a particular peer group, the locality-

aware peer 600 may initiate overlay network connections to peers that are members of that peer group. When establishing a new peer group, the locality-aware peer 600 may initiate overlay network connections to peers in existing peer groups that are nearby in the transport network. The establishment of new peer groups may be more common when the locality-aware overlay network is smaller, that is, when the locality-aware overlay network incorporates fewer locality-aware peers 600.

[0038] Figure 7 illustrates an example of a locality-aware overlay network 700 with multiple peer groups in accordance with an embodiment of the invention. Each peer group 702, 704, 706, 708, 710, 712, 714, 716, 718, 720 includes at least one locality-aware peer. The locality-aware overlay network 700 has a rendezvous point 722. Locality-aware peer 724 of Figure 7 has not yet joined overlay network 700.

[0039] Each peer group has at least one overlay network connection to another peer group, that is, at least one of the peers in each peer group has an overlay network connection to a peer in another peer group. In overlay network 700 each peer group has at least one connection with another peer group that is nearby in the transport network. For example, peer groups 704, 706 and 708 are relatively nearby peer group 702 in the transport network and peer group 702 has a direct overlay network connection to each of those peer groups, that is, peer groups 704, 706 and 708 are peer group neighbors of peer group 702. In contrast, peer group 718 is relatively distant from peer group 702 in the transport network and peer group 702 does not have a direct overlay network connection to peer group 718. Peer group 702 and peer group 718 are not neighbors in overlay network 700. In an embodiment of the invention, a peer group may have an overlay network connection to a random other peer

group, as well as nearby peer groups, in order to, for example, reduce the risk of overlay network partition due to localized transport network failure.

[0040] Peer groups including locality-aware peers 600 may be considered as nodes in a peer group network. Figure 8 is a schematic representation of network layers associated with a conventional overlay network. In Figure 8, conventional overlay network 802 is supported by transport network 804. Figure 9 is a suitable schematic representation of network layers associated with a locality-aware overlay network. A locality-aware overlay network 902 (e.g., the locality-aware overlay network 700) is supported by transport network 904 (which may be the same as transport network 804 of Figure 8). In an embodiment of the invention, as part of maintaining the locality-aware overlay network 902, the locality-aware overlay network 902 maintains a peer group network 906.

[0041] Each peer group may include a peer that is a leader of the peer group (a "peer group leader"). The peer establishing a new peer group is typically the first leader of the new peer group. Each locality-aware peer 600 in a peer group may become the leader of the peer group. In an embodiment of the invention, inter-group overlay network connections are between peer group leaders. For example, in Figure 5, peer 312 is the leader of peer group A, and peer 314 is the leader of the peer group B. The peer group network 906 of Figure 9 may be a network of those peers in the locality-aware overlay network 902 that are leaders of peer groups.

[0042] In an embodiment of the invention, each peer in a peer group is sufficiently close together in the transport network that a measurement of the transport network distance between a peer and the peer group leader is a good

approximation of the transport network distance between the peer and each peer in the peer group containing the peer group leader. This peer group coherency may enable a peer searching for a peer group in its transport network locality to rapidly traverse the locality-aware overlay network (e.g., from a random starting point) by querying the network's peer group leaders for transport network locality data. To further enhance locality-aware overlay network traversal, each peer group leader may maintain a list of neighboring peer groups (i.e., peer groups nearby in the transport network) and the transport network distances to those neighboring peer groups. For example, in Figure 7, the leader of peer group 702 maintains a list containing references to the leaders of peer groups 704, 706 and 708, as well as the measured transport network distance to those peer groups.

[0043] The locality-aware peer 600 searching for a peer group in its transport network locality, having found at least one peer group, may check if the peer group is in its locality by querying the candidate peer group (e.g., by querying the peer group leader) for references to the candidate peer group's neighboring peer groups as well as the transport network distances to those neighboring peer groups as measured from the candidate peer group. The peer 600 may then measure the transport network distances from itself to the peer groups neighboring the candidate peer group. If the transport network distances as measured by the peer 600 are the same (to within a tolerance, e.g., 10 milliseconds roundtrip time, or 1 transport network hop) as the transport network distances as supplied by the candidate peer group, then, in an embodiment of the invention, the peer 600 has found a peer group in its transport network locality. The peer groups in the peer group network

act as dynamic landmarks for the searching peer 600. The peer 600 may join the peer group by, for example, further querying the leader of the peer group for a list of peers in the peer group with which to establish overlay network connections. If the peer group neighbor distances as measured by the peer 600 do not match those supplied by the candidate peer group, then the peer 600 has a list of new candidate peer groups (i.e., the neighbors of the rejected candidate peer group) to check and may, for example, select the candidate peer group with the minimum measured transport network distance to try next.

[0044] Figure 10 depicts an example procedure performed by the locality-aware peer 724 (Figure 7) that employs this strategy when joining the locality-aware overlay network 700 in accordance with an embodiment of the invention. The join locality-aware overlay module 602 of Figure 6 may perform the procedure depicted by Figure 10. The locality-aware peer 724 joining the locality-aware overlay network 700 may obtain initial reference to at least one peer group in the locality-aware overlay network 700 in a number of ways. A computer user may input the transport network address of a peer group leader directly, for example, an Internet protocol (IP) address, a transport network domain name, or a uniform resource locator (URL). Alternatively, each locality-aware peer 724 may have reference to a rendezvous point (RP) (e.g., the rendezvous point 722) for the locality-aware overlay network 700 which may be obtained, for example, via DHCP or its overlay network analog. In the example depicted by Figure 10, a boot list of one or more candidate peer groups is obtained from the rendezvous point 722 for the locality-aware overlay network 700 at step 1002.

[0045] At step 1004, the peer 724 measures the transport network distance between itself and each of the candidate peer groups in the boot list. For example, the peer 724 may send an explicit Measure message to the leader of each peer group and, upon receiving a reply to the Measure message, calculate the round trip time. In an embodiment of the invention, the inter-group maintenance module 608 (Figure 6) is responsible for responding to Measure messages and the like. Other suitable methods of measuring transport network distance may be employed as will be appreciated by one of skill in the art. At step 1006, the peer group in the current set of candidate peer groups with the minimum measured transport network distance from the peer 724 is selected as the next peer group to check. If the peer 724 has an initial reference directly to a peer group in the overlay network (e.g., other than from the rendezvous point) then the peer 724 may bypass steps 1002, 1004 and 1006 and begin at step 1008.

[0046] At step 1008, the peer 724 queries the selected peer group for references to the selected peer group's neighboring peer groups. For example, the peer 724 may send the leader of the selected peer group an explicit Get Neighboring Peer Groups message and receive references to the neighboring peer groups in reply. For example, if the peer 724 sends the Get Neighboring Peer Groups message to the leader of peer group 702 then, in an embodiment of the invention, the peer 724 receives a list of the leaders of peer groups 704, 706 and 708 in reply. In an embodiment of the invention, the inter-group maintenance module 608 (Figure 6) is responsible for responding to Get Neighboring Peer Groups messages and the like.

[0047] At step 1010, the peer 724 queries the selected peer group for the transport network distances to the selected peer

group's neighboring peer groups as measured by the selected peer group. For example, the peer 724 may send the leader of the selected peer group an explicit Get Neighboring Peer Group Distances message and receive the transport network distance for each neighboring peer group in reply. For example, if the peer 724 sends the Get Neighboring Peer Group Distances message to the leader of peer group 702 then, in an embodiment of the invention, the peer receives the transport network distances from peer group 702 to peer group 704, from peer group 702 to peer group 706 and from peer group 702 to peer group 708 in reply. In an embodiment of the invention, the inter-group maintenance module 608 (Figure 6) is responsible for responding to Get Neighboring Peer Group Distances messages and the like.

[0048] In an embodiment of the invention, each peer group leader maintains, in its inter-group cache 610 (Figure 6), a list of peer group leaders in neighboring peer groups as well as the transport network distances to each neighboring peer group leader. Establishing and maintaining the list of neighboring peer groups is described below in more detail with reference to Figure 11. The inter-group maintenance module 608 may regularly update the transport network distances by, for example, regularly sending Measure messages to each peer group leader in the list. In an embodiment of the invention, requesting a copy of the inter-group cache 610 of the leader of the selected peer group provides at least the same information as the messages of steps 1008 and 1010 described above. Cache synchronization between peers of an overlay network may be supported by conventional overlay network protocols.

[0049] At step 1012, the peer 724 measures the transport network distances between itself and each of the neighboring peer groups of the selected peer group. For example, the peer

724 may send Measure messages to the leaders of each of the neighboring peer groups. For example, if the peer 724 is considering peer group 702 of Figure 7 then the peer 724 may send Measure messages to the leaders of peer groups 704, 706 and 708. At the completion of step 1012, the peer 724 has: a list of the neighboring peer groups of the selected peer group, the transport network distance from the selected peer group to each of those neighboring peer groups (the supplied distances), and the transport network distance from the peer to each of those neighboring peer groups (the measured distances).

[0050] At step 1014, the peer 724 compares each of the supplied distances to its corresponding measured distance. For example, if the peer 724 is considering peer group 702, the peer 724 first compares the transport network distance between peer group 702 and peer group 704, as supplied by peer group 702, to the transport network distance between the peer 724 and peer group 704, as measured by the peer 724. Next the peer 724 compares the transport network distance between peer group 702 and peer group 706, as supplied by peer group 702, to the transport network distance that the peer 724 measures between itself and peer group 706. Then the peer 724 compares the transport network distance between peer group 702 and peer group 708, as supplied by peer group 702, to the transport network distance between the peer and peer group 708, as measured by the peer 724. If, for each comparison, the absolute value of the difference (i.e., the value ignoring sign) is less than a threshold value (e.g., 10 milliseconds) then the selected peer group is determined to be in the transport network locality of the peer 724 and the procedure progresses to step 1016 where the peer 724 joins the selected peer group. Otherwise, the currently selected peer group is

determined not to be in the transport network locality of the peer 724 and the procedure progresses to step 1018.

[0051] At step 1018, the peer 724 examines the remaining set of candidate peer groups to determine if any are suitable for locality testing. It may be that all suitable candidates have already been tested or that too many (e.g., 10) peer groups have already been tested without finding a peer group in the peer's 724 locality. If there are no suitable candidate peer groups remaining, then the procedure progresses to step 1020 where the peer 724 establishes a new peer group with itself as the first member. If the set of candidate peer groups does still include suitable candidates then the procedure returns to step 1006 where the suitable candidate with the minimum measured transport network distance from the peer 724 is selected for the next round of locality testing.

[0052] For example, if the peer 724 is seeking to join overlay network 700 and the rendezvous point 722 provides peer group 702 as the starting point for the locality search then the peer 724 queries peer group 702 for the neighboring peer groups of peer group 702 and receives peer groups 704, 706 and 708 in reply. In an embodiment of the invention, these peer groups 704, 706 and 708 become the set of candidate peer groups for the next round of locality testing. If peer group 702 fails the locality test of step 1014 then the peer 724 selects the peer group in the set of candidate peer groups that is closest to the peer in the transport network as measured by the peer 724 in step 1012.

[0053] In this example, peer group 708 is the closest of the peer groups 704, 706 and 708. As a result, the peer 724 selects peer group 708 to be tested and queries peer group 708 for its neighbors. In overlay network 700, the neighboring

peer groups of peer group 708 are peer groups 702 and 710. In an embodiment of the invention, peer group 702 has already been tested and rejected, so it is not a suitable candidate, but peer group 710 is added to the set of candidate peer groups. The transport network distance to peer group 710 is measured as part of the procedure of testing peer group 708. If peer group 708 fails the locality test of step 1014 then, in an embodiment of the invention, the peer 724 selects the closest of peer groups 704, 706 and 710 to be tested next.

[0054] If peer group 710 is the closest, the peer group network search proceeds to peer group 710. In querying peer group 710 for its neighbors, peer groups 712, 714 and 716 are added to the set of candidate peer groups and so on until one of the peer groups passes the locality test, each of the peer groups in the locality-aware overlay network fails the locality test, or the peer 724 reaches a peer group network search limit in terms of time (e.g., 10 seconds), number of peer groups tested (e.g., 10) or the like. In an embodiment of the invention, peer groups already tested are not excluded from the set of candidate peer groups and if one of the peer groups that has already been tested is determined to be the best (i.e., minimum distance) candidate then the peer 724 establishes a new peer group with that peer group as the first neighbor of the new peer group.

[0055] In an embodiment of the invention, establishing a new peer group includes generating a new peer group identifier (ID) and establishing overlay network connections to a number of the new peer group's nearest neighbors. The nearest peer groups to the new peer group may be found with a variation of the method described above with reference to Figure 10. Having iterated through a peer group network and having decided to establish a

new peer group rather than join one of the existing peer groups, the peer 600 may have reference to the closest existing peer group (e.g., the last tested peer group). This closest existing peer group may become the new peer group's first neighbor. The new peer group may then query this first neighbor for its neighbors, measure their distances from the new peer group and select the closest of them for the new peer group's second neighbor. The new peer group may then query this second neighbor for its neighbors and so on until, for example, the new peer group has a suitable number of connected neighbors (e.g., 6 neighbors), the new peer group has connected to each of the peer groups in the locality-aware overlay network, or the peer group has connected to each peer group within a transport network distance boundary (e.g., 1000 milliseconds or 7 transport network hops).

[0056] Figure 11 depicts an example procedure for establishing a new peer group in a locality-aware overlay network in accordance with an embodiment of the invention. The join locality-aware overlay module 602 of Figure 6 may perform the procedure depicted by Figure 11. At step 1102, the peer 724 (Figure 7) establishing the new peer group generates a new peer group identifier with a method similar to the conventional generation of peer identifiers, for example, the peer group identifier may be a globally unique identifier (GUID) generated with methods known in the art. At step 1104, the peer 724 establishes an overlay network connection to the leader of the closest known existing peer group in the overlay network. For example, the peer 724 may have utilized the method described with reference to Figure 10 to traverse the locality-aware overlay network 700. Starting at peer group 702, the peer 724 may have considered each of peer groups 708, 710, 714, 718 and

720 in turn, but each peer group having failed the peer's 724 locality test, the peer 724 determines to establish a new peer group with itself as the first member. In this example, peer group 720 is the closest known existing peer group in the overlay network 700, so the peer 724 establishes its first overlay network connection to peer group 720, for example, to the leader of peer group 720.

[0057] At step 1106, the peer 724 checks if it has reached its target number of inter-group connections. The target number of inter-group connections may vary depending on the requirements of an application utilizing the locality-aware overlay network but, in an embodiment of the invention, it is at least 2 and often higher. If the target has been met then this aspect of establishing the new peer group is complete. Otherwise the procedure progresses to step 1108 to obtain candidate neighbors. If, at some time in the future, one or more of the established connections to neighboring peer groups is lost, the peer 724 may re-enter this procedure, for example, at step 1106, in order to restore the number of inter-group connections to the target number.

[0058] At step 1108, the peer 724 queries its new neighbor for a list of the new neighbor's neighbors in a manner similar to step 1008 of Figure 10. For example, if peer group 720 of Figure 7 is the new peer group's first neighbor then the new peer group queries peer group 720 for a list of its neighbors and receives a reference to peer group 718 and to the new peer group itself in reply. In an embodiment of the invention, these neighbors of the first neighboring peer group become candidate neighbors for the new peer group. In subsequent iterations, the results of neighbor queries may be added to the list of candidate neighbors. At step 1110, the peer 724

measures the transport network distance from itself to each of the new candidate neighbors in a manner similar to step 1012 of Figure 10. For example, having queried peer group 720 of Figure 7 for its neighbors, the peer 724 measures the transport network distance to peer group 718.

[0059] At step 1112, the peer 724 checks its list of candidate neighbors to determine if any are suitable candidates. It may be that the new peer group has established overlay network connections to each of the peer groups of the locality-aware overlay network without reaching its target number of inter-group connections. It may be that the new peer group has established peer group level connections to each of the peer groups within a transport network distance limit. If there are no suitable neighbor candidates then the peer 724 may exit the procedure depicted in Figure 11. Otherwise, the procedure returns to step 1104 where the peer 724 selects the closest of the candidate peer groups as a new neighbor for the new peer group. If the procedure exits without establishing the target number of inter-group connections, the leader of the new peer group may periodically re-enter this procedure, for example, at step 1104, to search for new neighbors.

[0060] In an embodiment of the invention, the peer that establishes a new peer group becomes the first leader of that peer group. The establishing peer may remain the leader of the peer group indefinitely, but if the establishing peer leaves the locality-aware overlay network (e.g., by disconnecting from the transport network) then, in an embodiment of the invention, another peer group member takes on the role. If there are no other peer group members then the peer group may cease to exist.

[0061] In an embodiment of the invention, the peer group leader establishes and maintains a leadership list. The leadership list may include a number of references to peers that are members of the peer group. The leadership list need not include each member of the peer group. In an embodiment of the invention, the leadership list is stored in the intra-group cache 606 (Figure 6) of the locality-aware peer 600. The peer establishing the peer group may add itself as the first member of the leadership list as part of the process of establishing the peer group.

[0062] When a new peer seeks to join the peer group, the new peer may query the peer group leader for a list of candidate peers with which to establish intra-group overlay network connections or, at least, to query for additional candidates if the candidate peers have reached their connection maximums. For example, the new peer may send a Get Peer Group Members message to the peer group leader and receive the current leadership list in reply. In an embodiment of the invention, the intra-group maintenance module 604 is responsible for responding to Get Peer Group Members messages and the like. If the overlay network incorporates a peer cache synchronization protocol, the new peer may obtain the list of candidate intra-group peers by synchronizing its intra-group cache 606 with the intra-group cache 606 of the peer group leader. If the leadership list includes less than a target number of peers (e.g., 10) then the new peer may be added to the leadership list as a result of querying the peer group leader for the list of candidate intra-group peers.

[0063] Each peer in the peer group may have a copy of the leadership list. If the current peer group leader leaves the peer group (e.g., leaves the overlay network) then, in an

embodiment of the invention, the peer that is next in the leadership list becomes the peer group leader. The new peer group leader may remove the old peer group leader from the leadership list and add a new peer to the bottom of the leadership list. In an embodiment of the invention, the peer group leader periodically (e.g., every 5 seconds) sends an Alive message to each peer in the peer group, for example, by utilizing conventional overlay network flooding-style message propagation within the peer group. If the Alive message is not received within a time period (e.g., 10 seconds) then, in an embodiment of the invention, the current group leader is determined to have left the peer group.

[0064] It may be that each of the peers in the leadership list leaves the peer group so that the peer group is left leaderless and without clear succession. In this situation, each peer remaining in the peer group may attempt to become the new peer group leader. For example, once the situation has been detected through lack of Alive messages, each peer may wait a random amount of time (e.g., 0 to 10 seconds) and then send a New Leader message if such a message has not already been received from another peer in the peer group. In an embodiment of the invention, the peer that sends out the New Leader message with the earliest timestamp is the new leader. A number of rounds may be required because two or more peers in the peer group may send out the New Leader message within a time period (e.g., 1 second). In an embodiment of the invention, the intra-group maintenance module 604 is responsible for the processing of Alive messages and the like.

[0065] Each peer in the peer group may have a copy of the list of peer group leaders in neighboring peer groups. The new peer group leader may re-establish the inter-group overlay

network connections to the leaders of neighboring peer groups in the list. Alternatively, the new peer group leader may perform the procedure discussed above with reference to Figure 11 to re-establish inter-group connections.

[0066] For example, with reference to Figure 5, peer 314 is the leader of the peer group in transport network locality 404. Peer 314 was the first member of the peer group and, as part of the process of establishing the new group, peer 314 added itself to the leadership list stored in its intra-group cache. When peer 308 joined the peer group, it queried peer 314 for a list of candidate intra-group peers. As a result, peer 314 added peer 308 to second place in the leadership list and returned itself as a candidate peer with which to establish an intra-group overlay network connection. Peer 308 did establish an overlay network connection with peer 314. Peer 304 was next to join the peer group and then peer 310. Each time a similar sequence of actions occurred. If peer 314 left the peer group, peer 308 would become the peer group leader and would re-establish the inter-group connection to peer 312.

[0067] All references, including publications, patent applications, and patents, cited herein are hereby incorporated by reference to the same extent as if each reference were individually and specifically indicated to be incorporated by reference and were set forth in its entirety herein.

[0068] The use of the terms "a" and "an" and "the" and similar referents in the context of describing the invention (especially in the context of the following claims) are to be construed to cover both the singular and the plural, unless otherwise indicated herein or clearly contradicted by context. The terms "comprising," "having," "including," and "containing" are to be construed as open-ended terms (i.e., meaning

"including, but not limited to,") unless otherwise noted. Recitation of ranges of values herein are merely intended to serve as a shorthand method of referring individually to each separate value falling within the range, unless otherwise indicated herein, and each separate value is incorporated into the specification as if it were individually recited herein. All methods described herein can be performed in any suitable order unless otherwise indicated herein or otherwise clearly contradicted by context. The use of any and all examples, or exemplary language (e.g., "such as") provided herein, is intended merely to better illuminate the invention and does not pose a limitation on the scope of the invention unless otherwise claimed. No language in the specification should be construed as indicating any non-claimed element as essential to the practice of the invention.

[0069] Preferred embodiments of this invention are described herein, including the best mode known to the inventors for carrying out the invention. Variations of those preferred embodiments may become apparent to those of ordinary skill in the art upon reading the foregoing description. The inventors expect skilled artisans to employ such variations as appropriate, and the inventors intend for the invention to be practiced otherwise than as specifically described herein. Accordingly, this invention includes all modifications and equivalents of the subject matter recited in the claims appended hereto as permitted by applicable law. Moreover, any combination of the above-described elements in all possible variations thereof is encompassed by the invention unless otherwise indicated herein or otherwise clearly contradicted by context.